Excess Mortality in the Age of COVID-19

Data Mining for Insights into Death

Julie Kirkpatrick  
 University of Colorado Boulder  
 Boulder Colorado USA  
 juki9272@colorado.edu

Alex Melnick  
 University of Colorado Boulder  
 Boulder Colorado USA  
 alme9138@colorado.edu

Kyle Tomlinson  
University of Colorado Boulder  
 Boulder Colorado USA  
 kyto3309@colorado.edu

ABSTRACT

One of the main challenges of understanding the COVID-19 pandemic is the lack of quality inputs for the number of cases and deaths experienced in a given population. At times, particularly early in the pandemic, testing resources were limited and many cases went unconfirmed. As the pandemic progressed, testing was not conducted universally, leaving an unknown number of COVID infections undiagnosed. At the same time, while deaths due to COVID are reported more reliably, this count fails to account for the second-order effects of the pandemic. Measuring excess mortality and using that as the basis for investigating the consequences and actions taken in response to the pandemic held promise for better illuminating some of the unknowns.

Some questions we investigated:

* Were excess mortality rates higher during the pandemic at a statistically significant level?
* Could relative excess mortality rates be used to rank the success of various countries at controlling the pandemic?
* How closely did reported COVID deaths match overall excess mortality?
* Are stringent public health measures related to excess mortality rates?
* In the absence of testing, can the number of cases be predicted based on other, more easily measured predictors?

We generated normalized excess mortality rates based on a publicly available mortality database using linear regression and normalization by population. Using these quantities in concert with information from COVID datasets, we were able to visualize a significant increase in excessive mortality in the majority of countries in the period from 2020 to 2022. Some countries, such as Taiwan and South Korea, outperformed other countries overall as they were able to keep their excess mortality within two standard deviations of normal levels. Total excess mortality globally outstripped reported COVID deaths, indicating that second-order mortality increased along with COVID or that deaths attributed to COVID did not fully account for all COVID deaths. Finally, we found a positive correlation between excess mortality and the stringency of public health measures. This correlation diminished in the weeks following the implementation of such measures, possibly indicating effectiveness over time. Attempts to correlate vaccination rates to overall mortality rates and attempts to predict cases based on other factors were inconclusive.

CCS CONCEPTS

• Applied computing~Life and medical sciences~Health care information systems

KEYWORDS

COVD-19, Mortality, Data Mining, Excess Mortality, Coronavirus, Pandemic, Data Collection

ACM Reference format:

Julie Kirkpatrick, Alex Melnick and Kyle Tomlinson. Excess Mortality in the Age of COVID-19: Data Mining for Insights into Death.

1  Introduction

When a global pandemic hits, there are many questions to answer and few clear answers. At the most basic level, there is debate in society over a question as simple as “Have people died as a result of the COVID-19 Pandemic?” The pandemic has disrupted society on myriad levels as governments and people make attempts to control the fallout from the pandemic. Using COVID datasets and a derivation of excess mortality, we have attempted to answer five key questions:

*Were excess mortality rates higher during the pandemic at a statistically significant level?*

Frustration at lockdowns, mask mandates, and other public health measures has some people questioning whether attempting to control the pandemic is even necessary. The ability to clearly show that overall mortality has risen significantly during the pandemic will provide a foundational reason for modifying personal behavior and implementing public health controls and counteract the claim that deaths attributed to COVID are merely relabeled deaths from other causes. A clear increase in overall mortality will indicate that the pandemic does, in fact, cause more deaths than the null hypothesis.

Could relative excess mortality rates be used to rank the success of various countries at controlling the pandemic?

Different countries approached the pandemic in different ways, such as the strict quarantine of New Zealand or the relative non-intervention of Sweden. Normalizing and comparing excess mortality may provide a method of measuring which national approaches and circumstances enjoyed the most success.

How closely did reported COVID deaths match overall excess mortality?

There are both challenges and incentives/disincentives to accurately reporting COVID deaths and measuring the toll of the pandemic. Assessing how closely the reporting of the world and individual countries matched the overall excess mortality will provide insight into where the deltas exist.

Are stringent public health measures related to excess mortality rates?

This question seeks to determine if public health measures do, in fact, lead to a reduction in mortality during the pandemic. If so, this will provide motivation for public health controls in future pandemics and insight into the calculus of weighing public health against competing interests.

In the absence of testing, can the number of cases be predicted based on other, more easily measured predictors?

Without testing the population regularly, it is difficult to gain a true understanding of the number of COVID cases during the pandemic, which in turn makes it difficult to assess the efficacy of public health measures or conditions that are correlated to pandemic outcomes. Being able to reliably reverse engineer case counts based on excess mortality or other factors might provide a more complete picture of the true case count.

2 Related Work

A substantial amount of research has been done in regards to the pandemic’s effects on mortality rates. Unfortunately, factors such as availability of testing supplies and political influences on public health messaging result in substantial differences between countries and subregions in reporting deaths as attributable to COVID-19 [2, 3, 6]. As a result of these differences public health research has often focused on a different metric: excess mortality, which refers to the number of deaths occurring over and above predicted mortality rates for a population [4]. The variation between reported COVID-19 mortality and excess mortality rates can be staggering: one example country reported 300,000 COVID-19 deaths but had over one million projected deaths over and above expected baseline mortality rates [6]. The majority of prior work focuses on either total excess mortality, a value that is relatively easier to state with confidence, or excess mortality directly attributable to COVID-19.

Predicted mortality rates establish a baseline upon which to compare mortality rates during the pandemic. The four-year period between 2015-2019 is commonly used to create these predicted rates[6, 4, 9], though some models opt to use different periods for a variety of reasons[2, 4, 5]. One of the primary challenges in creating an accurate baseline is the presence of non-pandemic-related factors that affect mortality rates. Changes in population demographics such as age can exhibit a powerful effect on actual mortality[1].

One article in particular, What has happened to non-COVID mortality during the pandemic?[8], does seek to address the same basic question that our project is focused on. While the scope of said article is restricted solely to the United Kingdom, it does provide a possible roadmap for our own inquiries. Another (admittedly pre-print) article presents a possibility for teasing out changes in specific non-COVID-19 mortality types by projecting current mortality rate subtype units in lieu of local units [7].

3 Data Set

We will be utilizing 3 data sets:

Human Mortality Database (HMD), which contains mortality data by week and age group for 38 countries Accessed at: <https://www.mortality.org/>, COVID-19 Dataset (COVID) which contains Global COVID Statistics by country and date Accessed at: <https://ourworldindata.org/explorers/coronavirus-data-explorer> , and COVID-19 Stringency Index which contains time series information on countries and the stringency of their policies regarding the pandemic. They also housed a separate dataset detailing the vaccination rates of various countries. Accessed at: <https://ourworldindata.org/covid-stringency-index>

Human Mortality Database

|  |  |  |
| --- | --- | --- |
| Name | Type | Description |
| CountryCode | Nominal String | Identifies country |
| Year | Ordinal Integer | Year in 1 year increments |
| Week | Ordinal Integer | Week from 1-52 from 1st full week of year |
| Sex | Nominal String | male, female, and combined |
| D0\_14 | Ordinal float | # of deaths age 0 to 14 (float due to transforming input data) |
| D15\_64 | Ordinal float | # of deaths age 15 to 64 (float due to transforming input data) |
| D65\_74 | Ordinal float | # of deaths age 65 to 74 (float due to transforming input data) |
| D75\_84 | Ordinal float | # of deaths age 75 to 84 (float due to transforming input data) |
| D85p | Ordinal float | # of deaths age 85+ |
| DTotal | Ordinal integer | Total # of deaths for all age groups |

COVID-19 Dataset

|  |  |  |
| --- | --- | --- |
| Name | Type | Description |
| iso\_code | Nominal String | Identifies country |
| Date | Ordinal String | Date in MM/DD/YYYY format |
| total\_cases | Ordinal Integer | Daily total of COVID cases; does not decrease with recovery or death |
| new\_cases | Ordinal Integer | Daily new cases reported |
| total\_deaths | Ordinal Integer | Daily total of COVID cases |
| new\_deaths | Ordinal Integer | Daily new deaths reported |
| total\_cases\_per\_million | Ordinal float | total\_cases per million |
| new\_cases\_per\_million | Ordinal float | New\_cases per million |
| hops\_patients | Ordinal Integer | Daily total of COVID cases hospitalized; does decrease with recovery or death |
| population | Ordinal integer | Population of country; does not change |

COVID-19 Stringency Index

|  |  |  |
| --- | --- | --- |
| Name | Type | Description |
| Entity | Nominal String | Identifies country full name |
| Code | Nominal String | Identifies country code |
| Day | Ordinal String | Date in MM/DD/YYYY format |
| stringency\_index | Ordinal Float | The index on any given day is calculated as the mean score of the nine metrics, 0 -100, with 100 being the strictest |

COVID-19 Vaccination Dataset

|  |  |  |
| --- | --- | --- |
| Name | Type | Description |
| location | Nominal String | Identifies country full name |
| iso\_code | Nominal String | Identifies country code |
| date | Ordinal String | Date in MM/DD/YYYY format |
| People\_fully\_  vaccinated | Ordinal integer | Total number of people who have received 2+ vaccination shots |

**5 Main Techniques Applied**

**5.1 Data Cleaning and Preprocessing**

Tools used: Python with Pandas and numpy libraries

COVID

* Retained only the following attributes: country, date, total\_cases, new\_cases, total\_deaths, new\_deaths, total\_cases\_per\_million, hosp\_patients, and population
* Dataset was resampled from daily to weekly data. “New” columns were summed and “Total” columns were retained as the first value in each week.
* “year” and “week” columns were extracted from the date column

HMD

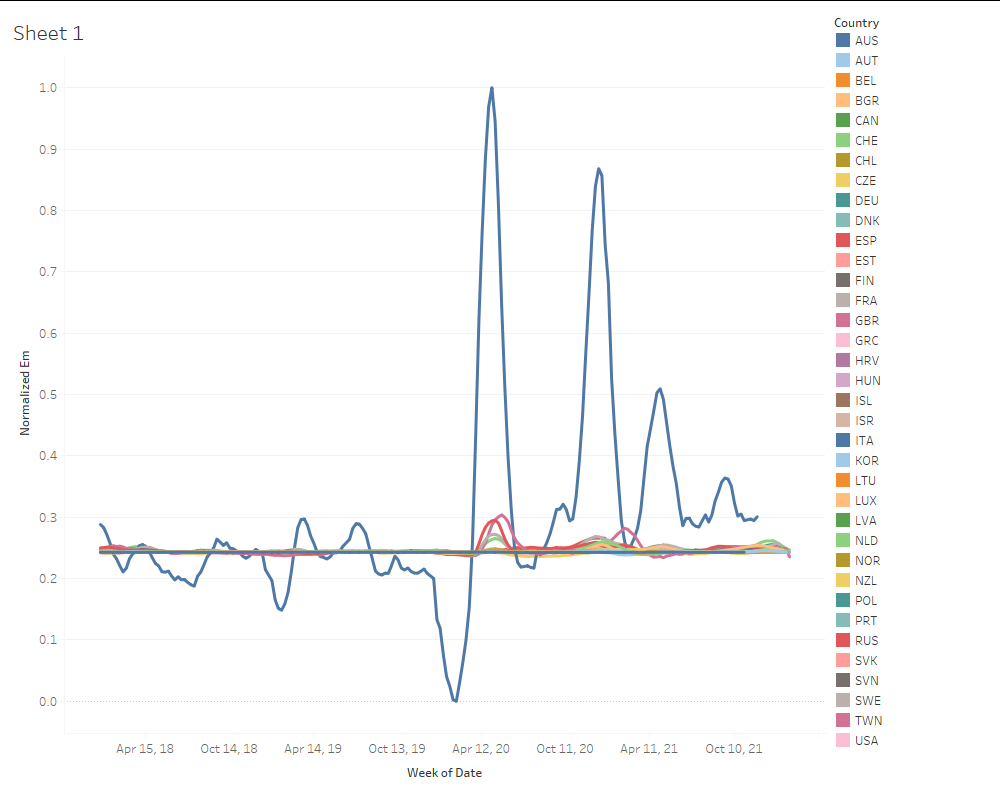
* The individual “male” and “female” categories were dropped while the combined category was retained
* Retained only the following attributes: country, week, year, DTotal
* Country codes that differed from the COVID dataset were transformed to match
* Great Britain data, split into England and Wales, Scotland, and Northern Ireland categories, was combined, summing the DTotal. Years prior to 2015 were eliminated due to missing information for Scotland and Northern Ireland in those ranges.
* “week” and “year” were transformed into a “date” column to assist in merging and visualizations

COVID Stringency Index

* Because we only have stringency index data from the beginning of the pandemic, another dataframe was made from the previously merged and cleaned datasets. This was done by filtering the Stringency Index data set first by the countries that we have excess mortality data on by excluding country codes that are not present in the cleaned, merged dataset. Then the day was converted to a datetime datatype, in order to pull the week number and year. The indices were then grouped by country code, week number, and year and averaged.

**5.2 Data Warehouse Techniques**

Though no formal data warehouse was established, two techniques were borrowed from the data warehouse scheme. First, aggregating the COVID datasets from daily data to weekly data was a form of drill-up. This allowed us to combine rows into a less granular form in order to match the HMD dataset. In addition, the aggregation had the effect of smoothing the data without a loss of fidelity. Secondly, use of Tableau as a tool allowed a much more efficient method of manipulating and filtering data from the datasets than could be accomplished by hand or through python, our principle tool for queries and manipulations. A good example of this was the initial plot of normalized excess mortality by country.



At a glance, Italy was such an outlier that it indicated some mistakes in the data processing. In fact, some population values had been misassigned while merging datasets, causing the false outliers. The efficiency of Tableau as a tool for interrogating the data allowed us to recognize and rectify this before drawing any erroneous conclusions.

**5.3 Data Mining**

Once both datasets were preprocessed and cleaned, we merged them into a single dataset. We then derived some additional attributes:

For both weekly deaths and weekly new cases, a moving average was computed by taking the mean of the nearest seven weeks in order to smooth the data and reduce the effects of random variation. Using the smoothed weekly death average for each individual week (e.g., week 1 of all years, week 2 of all years, etc.) we made a linear regression model based on ordinary least squares for each country to predict the number of expected deaths for each combination of week number and year. This was performed by week number to account for cyclical mortality rates that tend to rise and fall based on time of the year.

A linear regression was deemed appropriate as it would account for many unknown factors that make up the normal mortality rate for any given country, such as changes in healthcare standards, large disasters, or widespread external factors.

Subtracting the predicted weekly deaths from the actual deaths gave us an excess mortality number for each country/date combination.

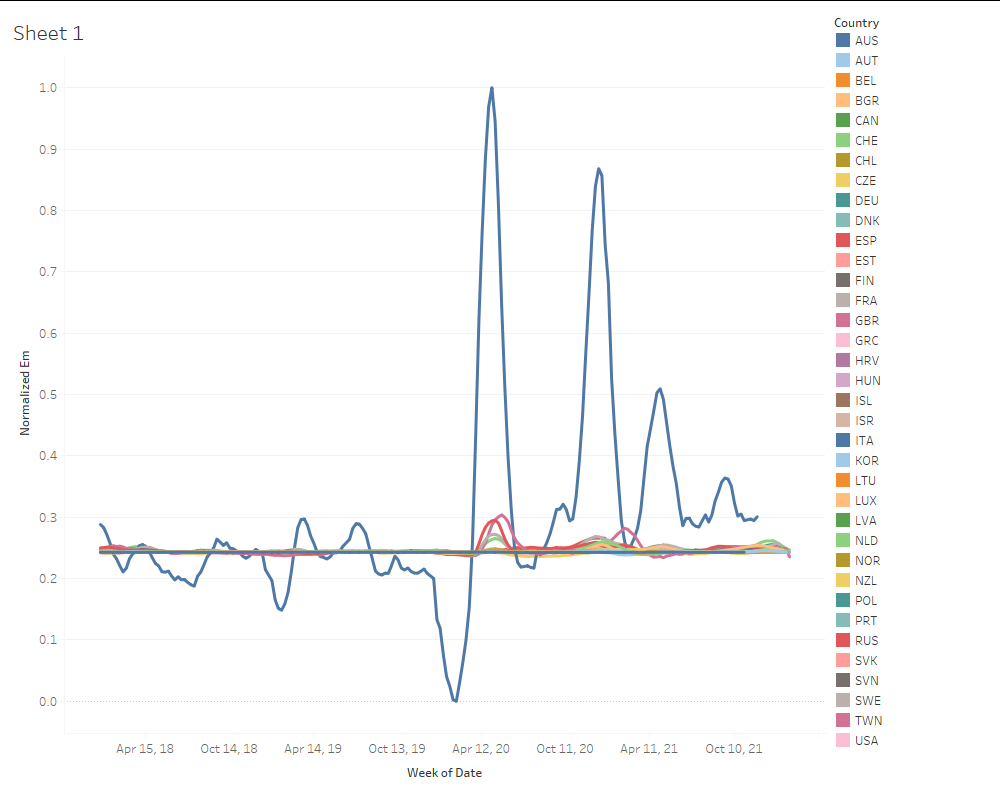
To normalize for population and better compare between countries, we divided excess mortality by population to create the em\_per\_capita attribute. We further normalized this measure by performing a min-max normalization across all countries.

To check these results, we placed the per capita excess mortality rates for all years prior to 2020 into a normal distribution using the statsmodel library. This allowed us to compute the standard deviation in order to evaluate the null hypothesis against the hypothesis that excess mortality rates had increased from 2020 to 2022 during the pandemic.

Visualizations were a powerful tool as we evaluated the significance of our work and interrogated the correlation between our data parameters.

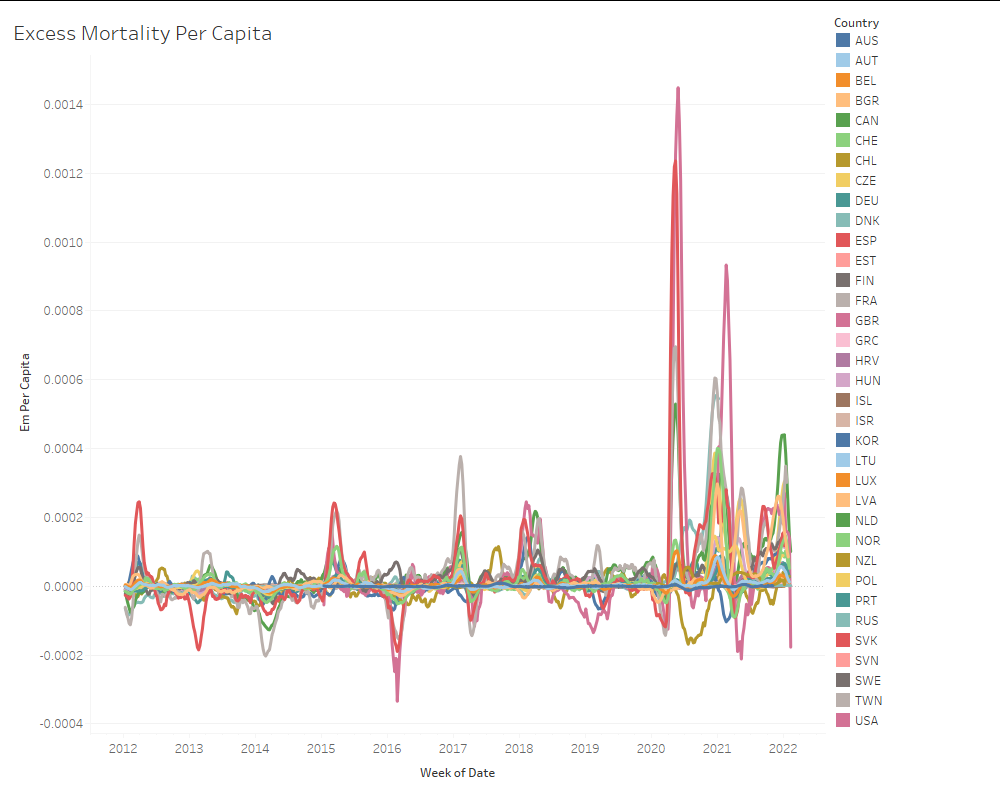
Stringency Index

Because we only have stringency index data from the beginning of the pandemic, another dataframe was made from the previously merged and cleaned datasets. This was done by filtering the Stringency Index data set first by the countries that we have excess mortality data on by excluding country codes that are not present in the cleaned, merged dataset. Then the day was converted to a datetime datatype, in order to pull the week number and year. The indices were then grouped by country code, week number, and year and averaged.

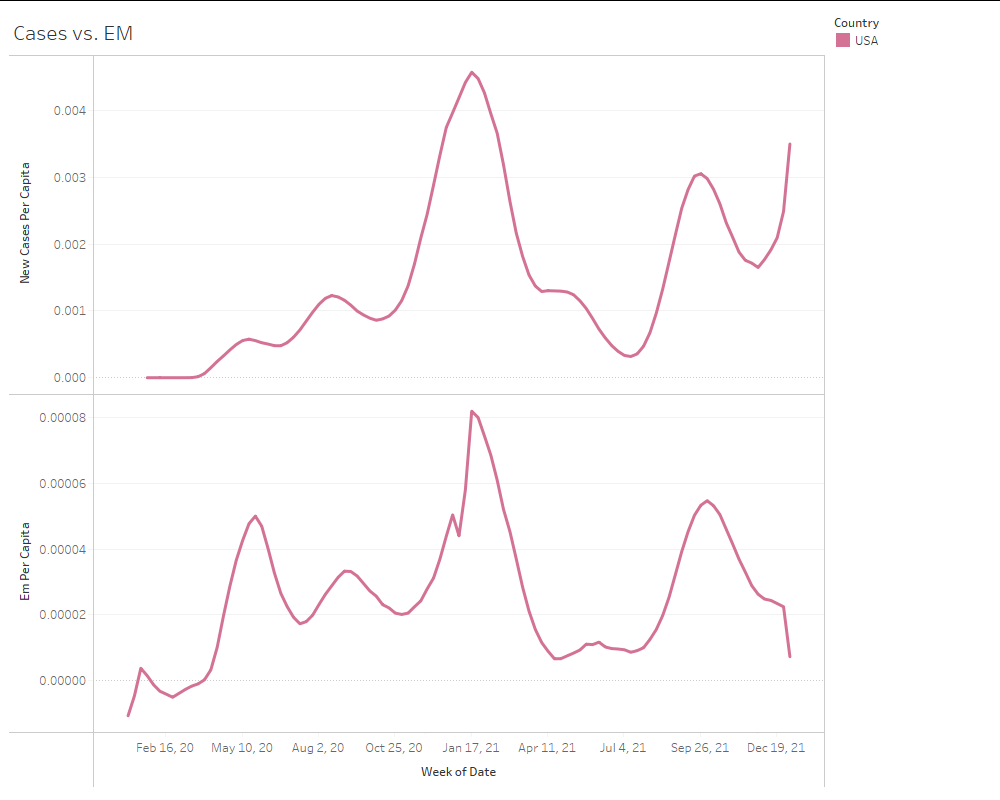


Visualizing the normalized excess mortality rates over time by country, it became clear that Italy’s mortality data was such an outlier that many other countries’ normalized data became indistinguishable between pre-COVID and post-COVID values. For the most part em\_per\_capita can stand in for the normalized values. For cases where normalized values will be useful, we created a normalization while dropping all Italian rows.

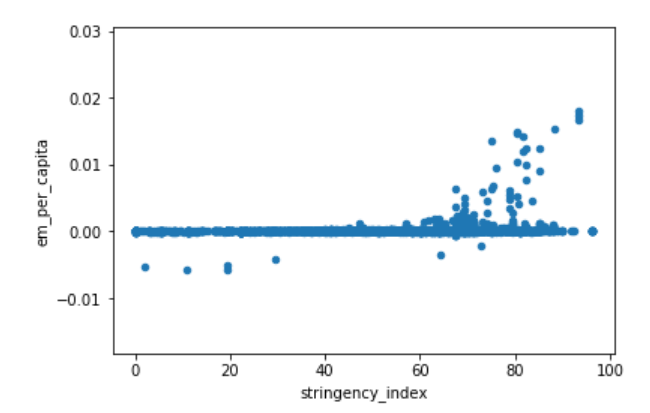
Initial exploratory data analysis using visualizations showed clear evidence of excess mortality during COVID, recreating portions of the past work in this area.



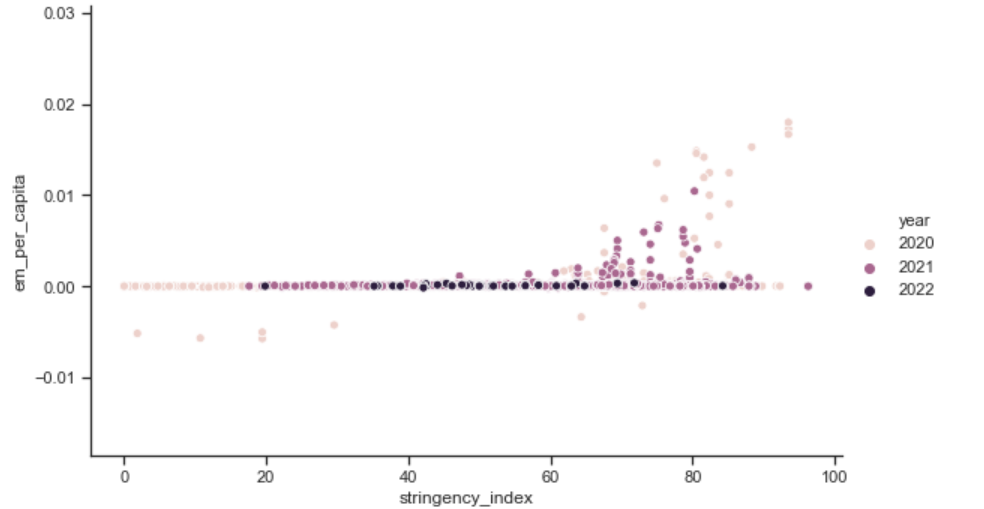
We can also see that excess mortality rates and COVID case rates appear to be correlated.



As for how government regulations affected excess mortality, from this scatterplot, we can see a slight correlation, with a large grouping past 60 on the stringency index.

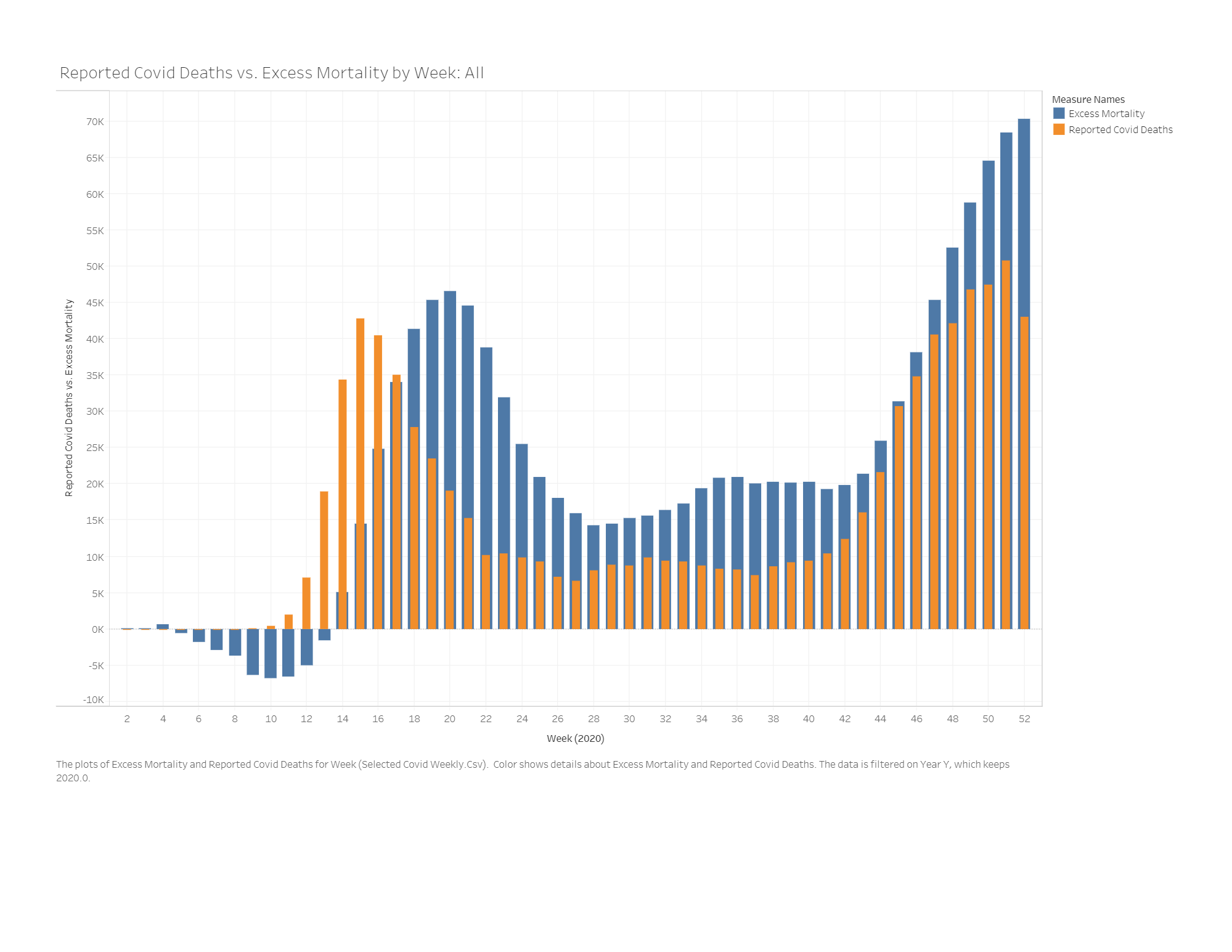


This leads us to examine the dimensionality of the set in a more meaningful way.



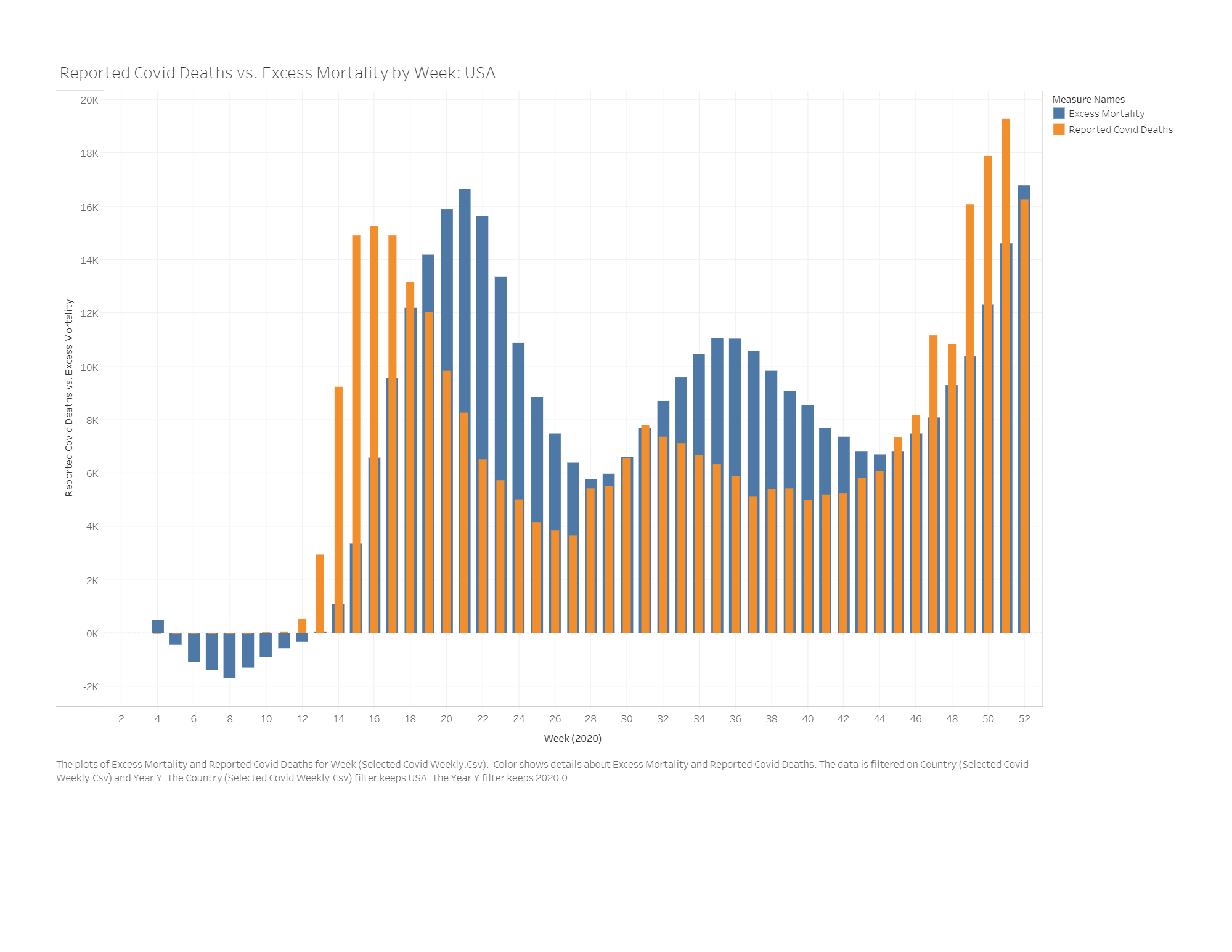
By coloring the points by year, we see the farthest reaching happened in year 1 of the pandemic. More filtering, processing, and analysis is needed.

We also have explored visualizations on reported COVID-19 Deaths and Excess Mortality for all countries.

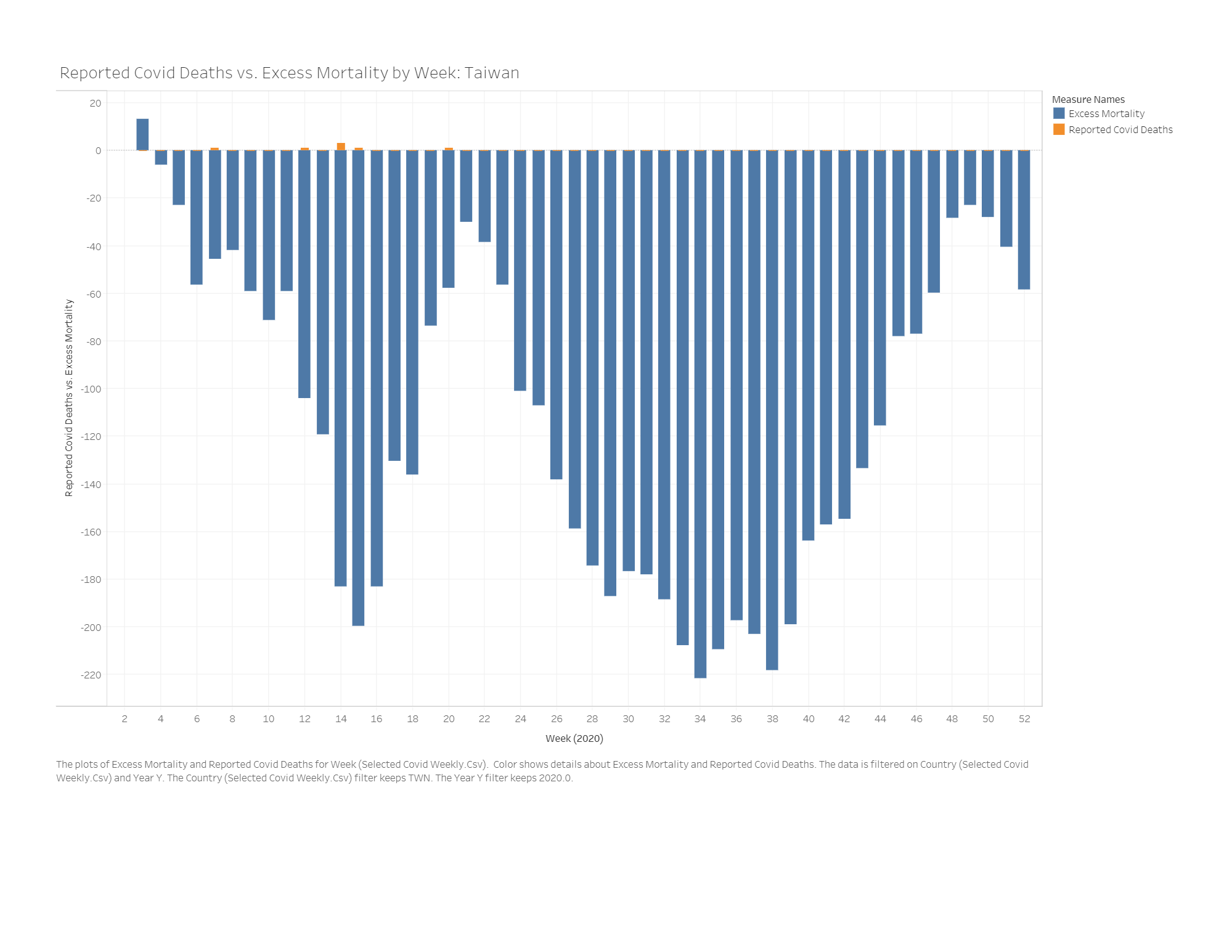


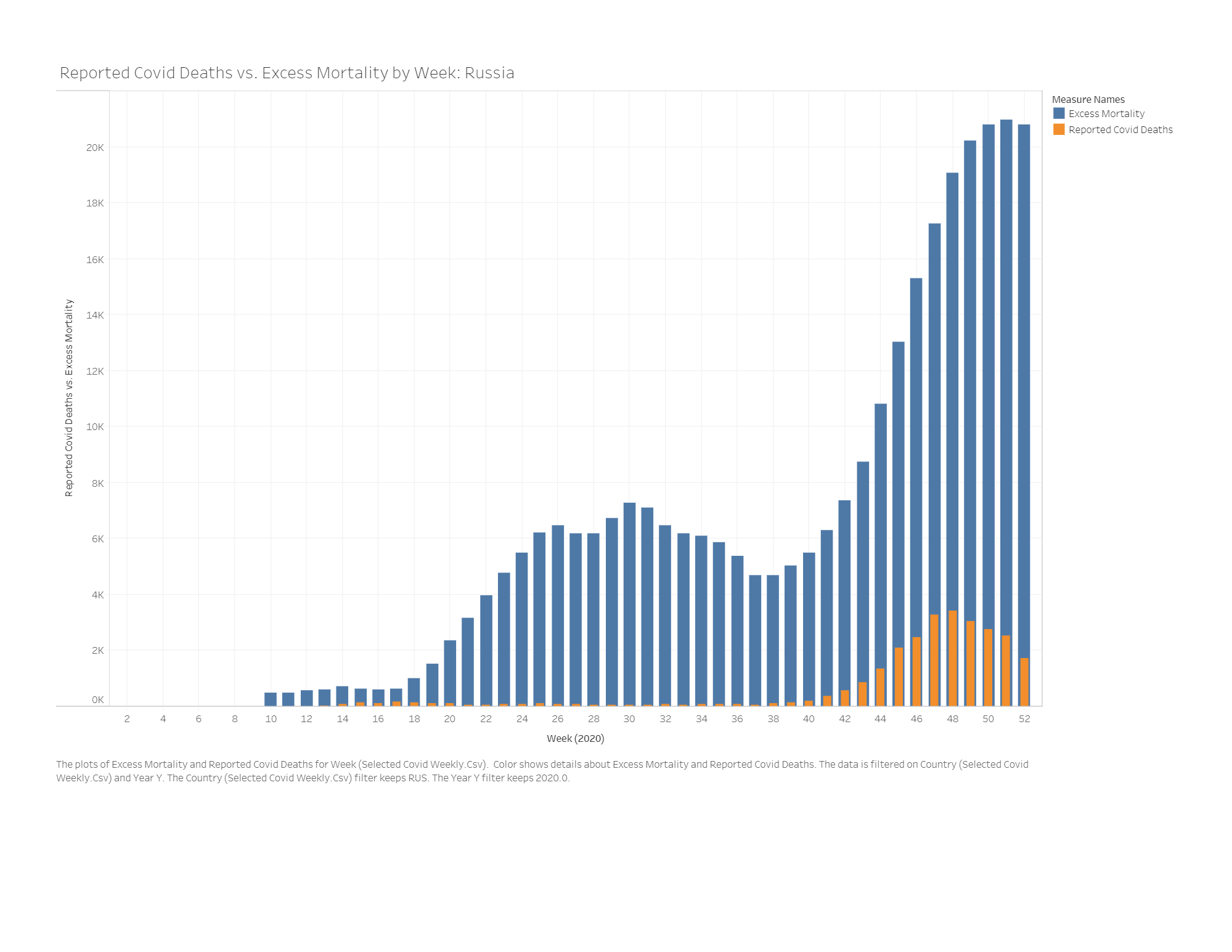
We also have visualizations for these attributes filtered by various countries:

United States:

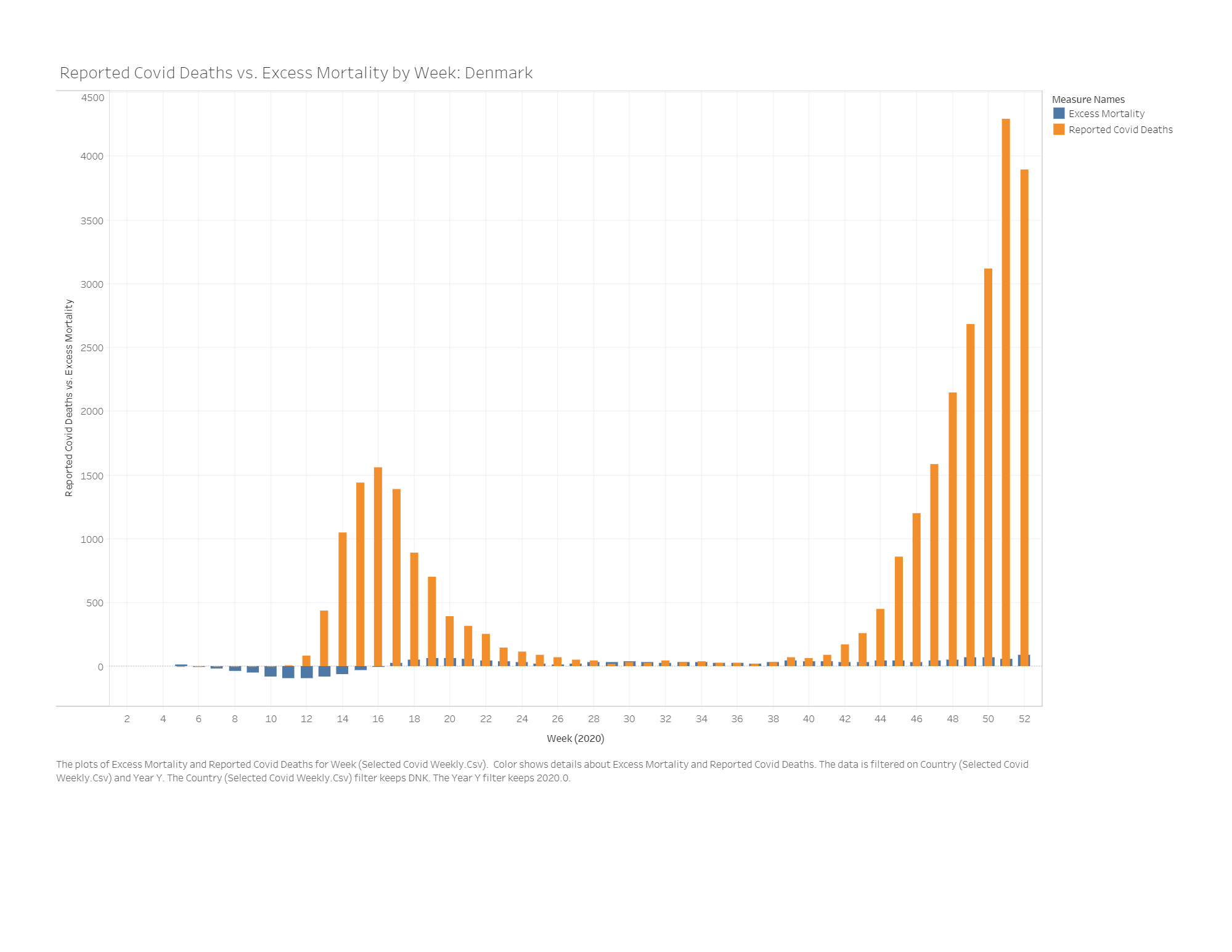


Taiwan:

Russia:



Denmark:



Clearly from the visualizations, there are many different scenarios that need further exploration into the explanation about why the excess mortality and reported COVID-19 deaths are or are not correlated.

REFERENCES

1. David Adam. 2021. The Effort to Count the Pandemic’s Death Toll. In *Nature 601, 312-315 (2022)*. <https://doi.org/10.1038/d41586-022-00104-8>
2. Abhishek Anand, Justin Sandefur, and Arvind Subramanian. 2021. Three New Estimates of India’s All-Cause Excess Mortality during the COVID-19 Pandemic. In *Center for Global Development, Working Paper 589, July 2021, Washington, DC*. <https://www.cgdev.org/publication/three-new-estimates-indias-all-cause-excess-mortality-during-covid-19-pandemic>
3. Héctor Pifarré i Arolas, Enrique Acosta, Guillem López-Casasnovas, Adeline Lo, Catia Nicodemo, Tim Riffe, and Mikko Myrskylä. 2021. Years of life lost to COVID-19 in 81 countries. In *Scientific Reports 11, 3504 (2021).* <https://doi.org/10.1038/s41598-021-83040-3>
4. Thomas Beaney, Jonathan M Clarke, Vageesh Jain, Amelia Kataria Golestaneh, Gemma Lyons, David Salman, and Azeem Majeed. 2020. Excess mortality: the gold standard in measuring the impact of COVID-19 worldwide? In *Journal of the Royal Society of Medicine. 2020;113(9):329-334*. <https://doi.org/10.1177/0141076820956802>
5. Nazrul Islam, Vladimir M Shkolnikov, Rolando J Acosta, Ilya Klimkin, Ichiro Kawachi, Rafael A Irizarry, Gianfranco Alicandro, Kamlesh Khunti, Tom Yates, Dmitri A Jdanov, Martin White, Sarah Lewington, and Ben Lacey. 2021. Excess deaths associated with covid-19 pandemic in 2020: age and sex disaggregated time series analysis in 29 high income countries. In *BMJ 2021;373:n1137*. <https://www.bmj.com/content/373/bmj.n1137>
6. Ariel Karlinsky and Dmitry Kobak. 2021. Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset. In *eLife 2021;10:e69336*. <https://doi.org/10.7554/eLife.69336>
7. Ariel Karlinsky. 2021. National Excess Mortality from Sub-National data: Method and Application for Argentina. Preprint in *medRxiv*. <https://doi.org/10.1101/2021.08.30.21262814>
8. Holly Krelle and Charles Tallack. 2021. *What has happened to non-COVID mortality during the pandemic?* The Health Foundation. <https://www.health.org.uk/publications/long-reads/what-has-happened-to-non-covid-mortality-during-the-pandemic>
9. Francesco Sanmarchi, Davide Golinelli, Jacopo Lenzi, Francesco Esposito, Angelo Capodici, Chiara Reno, and Dino Gibertoni. Exploring the Gap Between Excess Mortality and COVID-19 Deaths in 67 Countries. In *JAMA Network Open. 2021;4(7):e2117359*. <https://doi.org/10.1001/jamanetworkopen.2021.17359>

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00